Final Project: Group 5. Effect of Russian Trolls on the 2016 U.S. Presidential election.

Ryan Appel, Angel Arribas Lopez, Drew Breyer, Humza Kahn,
Devanshi Patel, Hanna Shin, Poonam Siyag

Scale of Problem

Existing Data

Describe the scale of the problem using existing data / opinions from expert reports or academic papers

> Tennessee-Knoxville study analyzed 770,005 tweets in English from known Russian troll accounts

Who is the target of the cyberthreat?

> The U.S.A. Presidential election for support of Donald Trump

Do we know how many targets are affected in the population? What is the proportion of the target group in the overall population of interest?

> Many estimates based on the posts and their reach. The portion of interest is those who saw ads AND changed their vote as a result.

Ideal Conditions

Describe the ideal data you would need to obtain a precise estimate of scale of the problem (proportion of the targeted units in the overall population of interest)

> How many people were directly influenced by Trolls I.e. were exposed to troll material and changed their vote as a result.

What are the major obstacles for collecting these data?

> Many significant problems. What does Reach mean when there are retweets? And what about bots? Hard to tell authentic reach from imitation reach. Also, how can you attribute the impact of trolls to someone's vote? Did their vote change for other reasons?

Do we know how many targets are affected in the population? What is the proportion of the target group in the overall population of interest?

> By one measure, 770,000 messages sent from known Russian troll accounts but reach is difficult to tell because of retweets and 43 million other election related tweets sent over same timeline could be difficult to attribute to Russia.

Impact of Cyber Threat

Existing Studies

Describe the major results from existing studies (academic papers or reports) regarding the impact of your cyber threat on your outcomes of interest

- Cross-Platform State Propaganda: Russian Trolls on Twitter and Youtube during the 2016 U.S. Presidential Election
  - Troll accounts were primarily trying to help increase support for Donald Trump and conservative candidates.
  - Some accounts were "agnostic" trying to inflame partisan divisions by supporting either side.
- Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign
  - Conservatives retweeted Russian troll more often producing 36x more tweets
  - Only about 4.9% of conservative users retweeting content were "bots"
- Describe the data used in these studies
  - 1 - 770,005 tweets from known Russian troll accounts
  - 2 - 43 million elections-related posts shared on Twitter by 5.7 million users (Sept 16-Nov 9 2016)
- How do authors justify that the relationship between a cyber threat and their outcomes of interest are causal? In other words, describe their research design (DiD, RD, IV, matching, naïve regression, etc)
  - Observational: Sorted tweets by their "mean" ideologies using latent semantic analysis
  - DiD through comparing Members of Congress "ideology" to discover categories of the tweets
- Describe in detail the credibility of these results: potential concerns with data (e.g., selection bias) and the research design (e.g., lack/improper control group, violation of parallel trends in DiD, self-sorting in RD, exclusion restriction in IV, etc)
  - The findings are credible given they are looking at a subset of the larger population of tweets that they've used strong methods to detect "troll-ness" but there is possibility of violation of parallel trends is certainly a difficult bar to meet. This requires the absense of treatment in the diffrences over-time which is something the authors did not address. This could potentially lead to biased estimation of the causal effect.

Ideal Experiment

- The observation can be voters who are in favor of either of Democratic or Republican party and do Twitter
- Treatment group will be exposed to foreign trolls through Twitter (retweet the foreign trolls)
- We can code treatment group as 1 if they retweet the foreign trolls
- Control group can be coded as 0 which means that they are not exposed to foreign trolls (do not retweet the foreign trolls)
- Our outcome of interest will be the change in voters' behavior after they are exposed to foreign trolls
- After the exposure to foreign trolls, voters' preferences change
- It is feasible. We can comprehend voters' preference based on how many the users have tweeted on the democratic party or the republican party. However, we need to know what twitter users are foreign trolls.
- We can do Difference in Difference design based on before and after exposure to foreign trolls

Existing Data Sources

- We can use dataset which Twitter provides. It has retweet information. We can find that who retweet the foreign trolls if we know which users are foreign trolls.
- Also, we can use Twitter dataset to know potential outcomes of interest. We can find the change in their preference based on post and retweet information such as what they post and what they retweet.

Structure of the Dataset

- Our structure of a dataset to estimate the impact of the foreign trolls
  - Vote preference 2012 election
  - Vote preference 2016 election
  - Exposure to foreign trolls
- Treatment variable
  - exposure to foreign trolls

- Outcomes of interest
  - Vote preference in 2016 election

Research Design

- Difference in Difference is part of “Design based inference” quasi experimental models.
- Widely used technique amongst economists, social scientists, and researchers.
- Top model to use for research better than observational studies
- Difference in difference method does not require randomization.

	2012	**2016**	**Difference**
Exposed to tweets (treatment)	A (not yet treated)	B (treated)	B - A
Not Exposed (control)	C (never treated)	D (never treated)	D - C
Difference	A - C	B - D	(B - A) - (D - C)

Code Analysis

TODO: replace the below tables with the actual R code to do the analysis and printing of the tables

Voting preference	pro democratic	anti democratic	pro republican	anti republican	troll tweets	vote 2012	vote 2016	vote changed	main information source
R	14	64	85	18	181	R	O	1	Online
R	21	72	90	23	206	R	R	0	Online
R	25	62	89	22	198	R	R	0	Social Media
R	8	73	96	29	206	R	R	0	Television
R	22	74	82	28	206	R	D	1	Online
R	23	60	85	18	186	R	R	0	Social Media
R	17	75	81	14	187	R	R	0	Social Media

Total Tweets	vote 2012	vote 2016	vote change	exposure to troll tweets
181	R	O	1	0
206	R	R	0	1
198	R	R	0	1
206	R	R	0	1
206	R	D	1	1
186	R	R	0	0

Regression Table Analysis

TODO: add the code to do the analysis as well as the above pictures missed in the previous section

Credibility

TODO: establish Credibility.

Conclusions

The size of the effect between voting behavior and those affected by troll tweet exposure is significant and shows evidence for our original hypothesis. Challenges in collecting actual data: the preference survey is a sensitive issue because it is political in nature which can cause a problems for data collection.

Policy to Mitigate a Cyber Threat

Existing Studies

Describe the major results from the existing studies (academic papers or reports) regarding the mitigation of your cyber threat

> Partnerships with social media platforms for Curbing social bots may be an effective strategy for mitigating the spread of online misinformation.

Describe the data used in these studies, Justify that the relationship between a policy and the scale of a cyber threat are causal, potential concerns with data

> There are no existing studies that can help us specify data for the mitigation of cyber threat.

Ideal Experiment

- We would partner with twitter to run an experiment on a subset of our current population. In this experiment, people would be exposed to fake tweets and legit tweets (all of them generated by us), and we would collect the following columns:
  - Number of fake troll tweets seen
  - Number of fake troll tweets reported
  - Number of fake troll tweets interacted with
  - Plan to vote for in 2024 (after the experiment period ends)
- The control group would be those people that see fake troll tweets but do not know whether they are troll tweets or not. They can interact with them but have no reporting capabilities nor learning material about the troll tweets.
- The treatment group would be those people that see fake troll tweets but can check whether they are legit tweets or troll tweet. They can learn overtime from experience as well as from resources provided from within twitter's platform such as information campaigns.
- We would use a difference in differences approach - how many people changed their voting preference in the control group vs. how many people changed their preference in the treatment group.
- The biggest concern is getting Twitter support for this. Everyone that is enrolled knows that they are part of an experiment. We would need to employ people to somehow generate the fake tweets or create a program that does it for us. It is probably unfeasible for Twitter to allow a 3rd party to conduct research about how their platform operates.

Existing Data Sources

- There are no existing sources to collect a dataset that could help us establish the impact of our policy since it is a new experiment that would need to be run first in order to collect and analyze the data and the impact of the policy on the cyber threat.
- The structure of our dataset will consist of these eight columns:
  - general_voting_preference
  - number_of_troll_tweets_seen
  - number_of_troll_tweets_interacted_with
  - number_of_troll_tweets_reported
  - voting_plan_2024
  - treated
  - voting_plan_different
  - affected
- Anyone who has the treated column as true, is part of the treatment group. This is regardless as to whether or not they reported any tweets or interacted with any tweets.
- The outcome of interest would be a DID model comparing whether or not someone was affected by the troll tweets seen (number_of_troll_tweets_seen > mean(troll tweets seen)) combined with whether or not the user is part of the treatment group or not.

- These two variables would describe as to whether the voting plan was different or not.

Research Design

- We used python to randomly create the control and treatment group as well as the results of the experiment. We used the DiD model to show that there is a difference between those who are part of the treatment and those who are not part of it.
- We randomly selected people from the population, and that we randomly showed them legitimate and fake troll tweets. The control group was always given the same chance to interact with them without any knowledge of the tweets being fake or real.

TODO: include charts

Regression Table

TODO: include charts

Credibility

- The data was collected as random users interacted with the fake trolls without the knowledge of the experiment proves the validity of our research design.
- Since these users came from the same population of affected users in the initial experiment. All we must prove here is that these users were randomly sampled and evenly distributed. We employed a random selection algorithm to determine whether users were part of the treatment group or not. Here are the results of the division of groups of users.

Conclusions

- Although this data was generated with the intension to show that this experiment would succeed, we believe the experiment could work and a real research experiment and real data are needed in order to come up to a solid conclusion.
- The main challenge we face in collecting this data is Twitter support to allow us to run the experiment.